

An Optimization Model for Constrained Discriminant Analysis and Numerical Experiments with Iris, Thyroid, and Heart Disease Datasets

Richard J. Gallagher, Ph.D., Department of Medical Informatics,
Eva K. Lee, Ph.D., Department of Industrial Engineering and Operations Research
Columbia University, New York, New York
David A. Patterson, Ph.D., Department of Mathematical Sciences
University of Montana, Missoula, Montana.

A nonlinear 0/1 mixed integer programming model is presented for a constrained discriminant analysis problem. The model enables controlling misclassification probabilities by placing restrictions on the numbers of misclassifications allowed among the training entities and incorporating a "reserved judgment" region to which entities whose classifications are difficult to determine may be allocated. A linearization of the model is given, and preliminary numerical results for two medical and one non medical domain are presented.

INTRODUCTION

A fundamental problem in discriminant analysis concerns the classification of an entity into one of G ($G \geq 2$) *a priori*, mutually exclusive groups based upon specific measurable features of the entity. Typically, a discriminant rule is formed from data collected on a sample of entities for which the group classifications are known. Then new entities, whose classifications are unknown, can be classified based on this rule. Such an approach has been applied frequently in medical diagnoses where often a definitive classification of a patient can be made only after exhaustive physical and clinical assessments, after surgery, or perhaps even after the patient dies and an autopsy is performed. Hence, tests based on relatively inexpensive and unobtrusive clinical and laboratory type observations are used to aid in a diagnosis [1].

Most work in discriminant analysis has focused on *forced* discrimination rules; that is, rules that definitely classify a given entity into one of the G *a priori* groups. A forced rule is characterized by a partition $\{R_1, \dots, R_G\}$ of the feature space, where an entity with feature vector x is classified as coming from group g if, and only if,

$x \in R_g$, $g = 1, \dots, G$. Forced discrimination is an effective approach for groups that are reasonably well-separated on the feature variables. However, if the groups are not well-separated, then applying a forced rule may result in a high number of misclassifications. In such cases, it may be desirable to form a discrimination rule that allows less specific classification decisions — or even non classification of some entities — in order to satisfy constraints on the misclassification probabilities. Quesenberry and Gessaman [2] proposed a general model whereby an entity may either be classified into some subset of the G groups (i.e., rule out membership in the remaining groups), or be placed in a "reserved judgment" category. An entity is considered misclassified only when it is assigned to a nonempty subset of groups not containing the true group of the entity. Discrimination rules of this type are referred to as *partial* or *constrained* discrimination rules. While such a general model is intuitively appealing, one disadvantage is that when $G \geq 3$ there is no obvious definition of optimality among any set of rules satisfying the constraints.

A simplified version of the above model involves only incorporating the reserved judgment category. Thus, an entity is either classified as coming from one of the G *a priori* groups, or it is placed in the reserved judgment category. A discrimination rule of this type, sometimes referred to as a rule with a "reject" option, is characterized by a partition $\{R_0, R_1, \dots, R_G\}$ of the feature space, where R_0 denotes the reserved judgment region. The reserved judgment option allows classification of an entity to be postponed until further information is available (i.e., information other than that associated with the k features on which the discrimination rule is based).

It is much easier to devise reasonable definitions of optimality for this model than for the general model. For example, whereas for the general model, maximizing the probability of correct classification would result in the useless rule of classifying every entity into the subset consisting of all the groups, here such an optimization strategy is meaningful.

Previous work on partial discriminant analysis has focused mainly on the two-group model. In this case the general model is equivalent to the simplified version, and numerical approaches for obtaining optimal discrimination rules have been proposed (e.g., see Anderson [3], Habbema, Hermans and Van Der Burgt [4], and Broffitt, Randles and Hogg [5]). McLachlan [6] summarizes the research on the two-group case.

For three or more groups, most work has been on rules of the more general type, where there is no clear definition of optimality. For instance, some ad hoc nonparametric approaches have been suggested by Quesenberry and Gesseman [2] and Ng and Randles [7]. However, very little work has been published on numerical techniques for constructing constrained rules for the simplified model for the three-or-more group case. Anderson [3] published an important result on the form of optimal rules of this type. However, to the authors' knowledge, no computational method based on Anderson's result has appeared in the literature.

In this paper we present an optimization model (specifically, a nonlinear 0/1 mixed integer program) based on Anderson's result that is applicable to the simplified model with any number of groups. The 0/1 indicator variables are used to represent if a training entity is assigned to a given region, and the model seeks to maximize the number of correct classifications of the training entities while placing upper bounds on the numbers of misclassifications. A heuristic linearization for the nonlinear model is suggested, and preliminary numerical results for the linearized model, using two medical and one non medical dataset obtained from the Machine Learning Database Repository at the University of California at Irvine [8], are presented.

OPTIMIZATION MODEL

Under quite general assumptions, Anderson [3] showed that an optimal partition for a discriminant rule with a reject option, where one is seeking to maximize the probability of correct classification

while maintaining specified bounds on the misclassification probabilities, is of the form

$$R_g = \{x : L_g(x) = \max_{h \in \{0,1,\dots,G\}} L_h(x)\},$$

for $g = 0, \dots, G$. Here, L_0 is the function that is identically zero, and the functions L_h , $h = 1, \dots, G$, are of the form

$$L_h(x) = \pi_h f_h(x) - \sum_{i \neq h} \lambda_{ih} f_i(x),$$

where f_h , $h = 1, \dots, G$, are appropriate group conditional density functions; π_h , $h = 1, \dots, G$, are prior probabilities for the groups; and the λ_{ih} 's are nonnegative parameters to be determined.

Assume now that we are given a training sample of N entities whose group classifications are known, and for which measurements on k feature characteristics have been made. Say n_g of the training entities are in group g , where $\sum_{g=1}^G n_g = N$; and let the k -dimensional vectors x^{gj} , $g = 1, \dots, G$, $j = 1, \dots, n_g$, contain the measurements on the k characteristics. Using this data, one can compute estimates \hat{f}_h for the group conditional density functions f_h , $h = 1, \dots, G$ (e.g., see [6]). Also, estimates $\hat{\pi}_h$ of the prior probabilities π_h , $h = 1, \dots, G$, must be made. Once these estimates are made, an appropriate set of λ_{ih} 's can be obtained by solving the nonlinear 0/1 mixed integer program (MIP) given below.

In the MIP model, as a surrogate to maximizing the probability of correct classification, the objective is to maximize the number of correct classifications of the given N training entities. Similarly, the constraints on the misclassification probabilities are modeled by ensuring that the number of group g training entities in region R_h ($h \neq g$) is less than or equal to a pre-specified percentage, p_{hg} ($0 < p_{hg} < 1$), of the total number, n_g , of group g entities. For notational convenience, let $\mathcal{G} = \{1, \dots, G\}$, and $\mathcal{N}_g = \{1, \dots, n_g\}$ for $g \in \mathcal{G}$. Then the mixed integer program can be written as:

$$\begin{aligned} & \text{maximize} \quad \sum_{g \in \mathcal{G}} \sum_{j \in \mathcal{N}_g} u_{ggj} \\ & \text{subject to} \\ & L_{hgj} = \hat{\pi}_h \hat{f}_h(x^{gj}) - \sum_{\substack{i \in \mathcal{G} \\ i \neq h}} \lambda_{ih} \hat{f}_i(x^{gj}) \end{aligned} \quad (1)$$

$$y_{gj} = \max\{0, L_{h_{gj}} : h = 1, \dots, G\} \quad (2)$$

$$y_{gj} - L_{ggj} \leq M(1 - u_{ggj}) \quad (3)$$

$$y_{gj} - L_{h_{gj}} \geq \epsilon(1 - u_{h_{gj}}) \quad (4)$$

$$\sum_{j \in \mathcal{N}_g} u_{h_{gj}} \leq p_{hg} n_g \quad (5)$$

$$y_{gj} \geq 0, \lambda_{ih} \geq 0, u_{h_{gj}} \in \{0, 1\}. \quad (6)$$

The continuous variables $L_{h_{gj}}$ and y_{gj} , and constraints (1) and (2) capture the essence of a discretized version of Anderson's result. In particular, $L_{h_{gj}}$ represents the value of the function L_h when evaluated at the training point x^{gj} , and y_{gj} represents the maximum of $\{L_h(x^{gj}) : h \in \{0, 1, \dots, G\}\}$. The 0/1 variables $u_{h_{gj}}$ are used to indicate whether or not x^{gj} lies in region R_h ; i.e., whether or not the j th entity from group g is allocated to group h . In particular, constraints (3), together with the objective, force u_{ggj} to be 1 if, and only if, the j th entity from group g is correctly allocated to group g ; and constraints (4) and (5) ensure that at most $p_{hg} n_g$ group g entities are allocated to group h , $h \neq g$. Though the parameters M and ϵ are extraneous to the discriminant analysis problem itself, they are needed in the model in order to control the indicator variables $u_{h_{gj}}$. The intention is for M and ϵ to be, respectively, large and small positive constants.

The nonlinearity of constraint (2) makes it impractical to solve the above optimization model directly. Although there are a variety of commercially available optimization solvers that can solve mixed integer *linear* programs, none have the capability of directly dealing with nonlinear constraints of the form (2). One heuristic approach to linearizing the model is to replace constraint (2) with the constraints $y_{gj} \geq L_{h_{gj}}$, $h = 1, \dots, G$, and include penalty terms in the objective function. In particular, our linearized model has the objective

$$\text{maximize} \quad \sum_{g \in \mathcal{G}} \sum_{j \in \mathcal{N}_g} \alpha u_{ggj} - \sum_{g \in \mathcal{G}} \sum_{j \in \mathcal{N}_g} \beta y_{gj},$$

where α and β are positive constants. This linearized model is heuristic in that there is nothing to force $y_{gj} = \max\{0, L_{h_{gj}} : h = 1, \dots, G\}$. However, since in addition to trying to force as many u_{ggj} 's to one as possible, the objective also tries to make the y_{gj} 's as small as possible, there is a tendency for the optimizer to drive y_{gj} towards $\max\{0, L_{h_{gj}} : h = 1, \dots, G\}$. We remark that the α 's and β 's could be stratified by group (i.e., introduce possibly distinct $\alpha_g, \beta_g, g \in \mathcal{G}$) to model the

relative importance of certain groups to be correctly classified. In our numerical tests, we set the weights to be the same for every group, assuming equal importance of each group.

NUMERICAL RESULTS

Ten-fold cross-validation was performed on datasets obtained from the machine learning database repository at the University of California at Irvine [8]. In ten-fold cross-validation, a dataset is randomly partitioned into ten subsets of equal size. Ten trials are then run, each of which involves a training set made up of nine of the subsets and a test set made up of the remaining subset. The classification rule obtained via a given training set is applied to the associated test set to determine the number of test points correctly classified, the number misclassified, and (in the present study) the number allocated to the reserved judgment region. Averages are computed from the ten trials to obtain unbiased estimates of the expected percentages of each of these three possible outcomes.

We focused on three datasets from the UCI repository: *iris*, *new-thyroid*, and *heart-disease*. *Iris* is a classic dataset used to test discrimination techniques. The data consists of measurements of the sepal length and width and petal length and width of fifty plants for each of three types of iris. Using all four measurements, the three groups are reasonably well-separated. However, when only the measurements on the sepal length and width are used, there is more overlap among the groups, and as such the dataset is a good test bed for a partial discrimination technique.

New-thyroid consists of data used in trying to predict the state of the thyroid gland. Three diagnostic classes are specified: euthyroidism, hypothyroidism and hyperthyroidism. Five laboratory measurements for 215 patients are provided: total serum thyroxine, as measured by the isotopic displacement method; total serum triiodothyronine, as measured by radioimmunoassay; T3-resin uptake (a percentage); basal thyroid-stimulating hormone (TSH), as measured by radioimmunoassay; and the maximal absolute difference of TSH after injection of 200 micrograms of thyrotropin-releasing hormone, as compared to the basal value. Coomans, Broeckaert, Jonkheer, and Massart [9] used this data in a study comparing sixteen different forced discrimination techniques. However, their study only focused on two-group discrimination problems (i.e., euthyroidism vs. hypothyroidism, and euthyroidism vs. hyperthyroidism), so the results are not directly comparable to the

results herein.

The *heart-disease* database consists of data pertaining to angiographic coronary disease. There are five diagnostic categories, ranging from 0 to 4, graded by the percentage of the narrowing of the diameter of a major blood vessel. Although there are 76 raw attributes, in published experiments to date [10, 11], only 13 of the 76 attributes have been used as predictive variables in deriving discriminant rules. These include age, sex, chest pain type, systolic blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST segment depression induced by exercise relative to rest, slope of the peak exercise ST segment, and the number of major vessels colored by fluoroscopy for coronary calcium. Moreover, efforts have concentrated on simply attempting to distinguish between presence (diagnostic values 1,2,3,4) and absence (diagnostic value 0) of disease. Although data was collected at four sites (the Cleveland Clinic Foundation; the Hungarian Institute of Cardiology; the V.A. Medical Center in Long Beach, California; and the University Hospital in Zurich, Switzerland), there is a considerable amount of missing data from all sites but the Cleveland Clinic Foundation, and only data collected at this site have been utilized for discrimination purposes in past studies. Likewise, our numerical experiments only utilize the Cleveland database. Unlike previous uses, however, we attempt to discriminate between all of the five diagnostic categories.

Preliminary numerical tests for the proposed discrimination technique are reported below. The goal of these preliminary tests is simply to obtain raw measurements on the performance of the technique. Further numerical work is currently underway to compare the technique with forced discrimination procedures, as well as with an alternative linearization of the nonlinear model.

Our tests were carried out with the allowed misclassification levels set at 15% ($p_{hg} = .15$); the extraneous parameters M and ϵ set at 100 and 0.01, respectively; and objective function weights set to $\alpha = 2$ and $\beta = 3$. Also, we used normal-model group-conditional densities (e.g., see [6]) and equal prior probabilities for all groups.

It should be noted that, although there are commercially available optimization software packages capable of solving mixed integer linear programs (e.g., [12]), such problems are, in general, very difficult to solve. Because of its availability to us, we used a state-of-the-art research code [13] when

running our ten-fold tests. However, as is common when applying optimization solvers to large-scale real world MIP instances, we terminated the solver after the first feasible solution satisfying constraints (1)–(6) was obtained. This typically occurred within one second of CPU time running on a single-processor Sun Sparc 20.

Once an initial feasible solution is obtained, the associated λ_{ih} 's are used to define the functions L_h , $h = 1, \dots, G$, which in turn are used to test membership of the test set entities in the regions R_g , $g = 1, \dots, G$. Although the current values of the 0/1 indicator variables accurately reflect the membership of most of the training entities, the heuristic nature of the model leaves open the possibility that some of the 0/1 variables may have the wrong value. Hence, we also test the training set entities against the functions L_h , $h = 1, \dots, G$ to determine to which region each entity is assigned.

The results of our experiments are summarized in Tables 1 and 2 below. For each dataset, we record the percentage of entities correctly classified, the percentage classified in the reserved judgment region, and the percentage misclassified, as averaged over the ten trials. We include results for both *iris* (with all four feature measurements used for discrimination purposes) and *sepal* (the iris data, when only sepal length and width are used) to observe how the model performs on both a known well-separated dataset and on one for which the groups are known to be mixed.

Table 1. Training set results

Dataset	Correct	Reserved	Misclass.
Iris	98.5	0.2	1.3
Sepal	79.1	7.2	13.6
New-thyroid	85.5	14.3	0.2
Heart-disease	70.3	18.2	11.5

Table 2. Test set results

Dataset	Correct	Reserved	Misclass.
Iris	95.3	1.3	3.3
Sepal	66.1	16.1	17.9
New-thyroid	81.8	17.3	0.9
Heart-disease	42.5	25.0	32.5

In light of the fact that the *iris* data is known to be well-separated, the results obtained for this dataset are what one would hope for: very high correct classification rates for both the training set and the test set, and very little misclassification. The reserved judgment region had little consequence in this case. Indeed, for a dataset that is fairly-well separated, a forced rule is adequate.

The results for *sepal* offer evidence that the proposed optimization model holds promise as a viable approach to constrained discrimination for cases when the groups are mixed. The technique

performed quite well on the training data, though less so on the test data.

New-thyroid has a high rate of correct classification and a low rate of misclassification for both the training set and the test set. Moreover, the reserved judgment category captures a significant portion of the entities.

The results for *heart-disease* are mixed. Although the training set results are reasonably good, there is a high misclassification rate on the test set data. This might be indicative of "over training." With 13 feature variables used on a dataset consisting of approximately 300 patients, and for which the smallest group (diagnostic group 4) has only 13 patients, there could be a tendency for the discriminant rules to over fit the training data, and thereby be less accurate with new data.

Also, it is important to emphasize again that the optimizer was halted after the first feasible solution was found. Consequently, there is no guarantee that this solution is optimal. Although one would expect that a true optimal solution to the MIP would lead to a good approximation to the optimal partition $\{R_0, \dots, R_G\}$, the partition defined by a suboptimal solution to the MIP may very well be far from optimal. This observation may contribute to the relatively poor performance of the discriminant rules for *heart-disease*, and to a lesser extent for *sepal*. In order to determine if this is indeed the case, the optimizer must be allowed to run to optimality so that a direct comparison can be made between the two partitions. Numerical work in this direction is currently being conducted.

CONCLUSION

With the potential serious consequences of misdiagnoses, developing discrimination rules that incorporate a reserved judgment region is an avenue worth investigating. The optimization model suggested herein is well founded on theoretical grounds, and the preliminary numerical work shows that it holds promise for being a viable approach to constrained discrimination when three or more groups are involved.

Acknowledgment

The first author was supported in part by grant number LM07079 from the National Library of Medicine. The second author was supported in part by NSF CAREER grant CCR-9501584.

References

1. Wagner G, Tautu P, Wolbler U. Problems of medical diagnosis – a bibliography. *Methods of Information in Medicine*, 1978:17:55-74.
2. Quesenberry CP, Gessaman MP. Nonparametric discrimination using tolerance regions. *Annals of Mathematical Statistics*, 1968:39:664-673.
3. Anderson JA. Constrained discrimination between k populations. *Journal of the Royal Statistical Society, Series B*, 1969:31:123-139.
4. Habbema JDF, Hermans J, Van Der Burgt AT. Cases of doubt in allocation problems. *Biometrika*, 1974:61:313-324.
5. Broffit JD, Randles RH, Hogg RV. Distribution-free partial discriminant analysis. *Journal of the American Statistical Association*, 1976:71:934-939.
6. McLachlan GJ. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.
7. Ng T-H, Randles RH. Distribution-free partial discrimination procedures. *Computers and Mathematics with Applications*, 1986:12A:225-234.
8. Murphy PM, Aha DW. *UCI Repository of machine learning databases* [<http://www.ics.uci.edu/~mllearn/MLRepository.html>], Department of Information and Computer Science, University of California, Irvine, California, 1994.
9. Coomans D, Broeckaert I, Massart DL. Comparison of multivariate discrimination techniques for clinical data — application to the thyroid functional state. *Methods of Information in Medicine*, 1983:22(2):93-101.
10. Detrano R, Janosi A, Steinbrunn W, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 1989:64:304-310.
11. Gennari JH, Langley P, Fisher D. Models of information concept formation. *Artificial Intelligence*, 1989:40:11-61.
12. Using the CPLEX Callable Library and the CPLEX Mixed Integer Library. Cplex Optimization, Inc., Incline Village, Nevada, 1993.
13. Lee EK, Bixby R, Cook W, Cox AL. Parallel mixed integer programming. Center for Research on Parallel Computation Research Monograph CRPC-TR95554, 1995.